Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development

Year: 2021 | Citations: 350 | Authors: Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf H. Roohani

---

Abstract

Therapeutics machine learning is an emerging field with incredible opportunities for innovatiaon and impact. However, advancement in this field requires formulation of meaningful learning tasks and careful curation of datasets. Here, we introduce Therapeutics Data Commons (TDC), the first unifying platform to systematically access and evaluate machine learning across the entire range of therapeutics. To date, TDC includes 66 AI-ready datasets spread across 22 learning tasks and spanning the discovery and development of safe and effective medicines. TDC also provides an ecosystem of tools and community resources, including 33 data functions and types of meaningful data splits, 23 strategies for systematic model evaluation, 17 molecule generation oracles, and 29 public leaderboards. All resources are integrated and accessible via an open Python library. We carry out extensive experiments on selected datasets, demonstrating that even the strongest algorithms fall short of solving key therapeutics challenges, including real dataset distributional shifts, multi-scale modeling of heterogeneous data, and robust generalization to novel data points. We envision that TDC can facilitate algorithmic and scientific advances and considerably accelerate machine-learning model development, validation and transition into biomedical and clinical implementation. TDC is an open-science initiative available at https://tdcommons.ai.